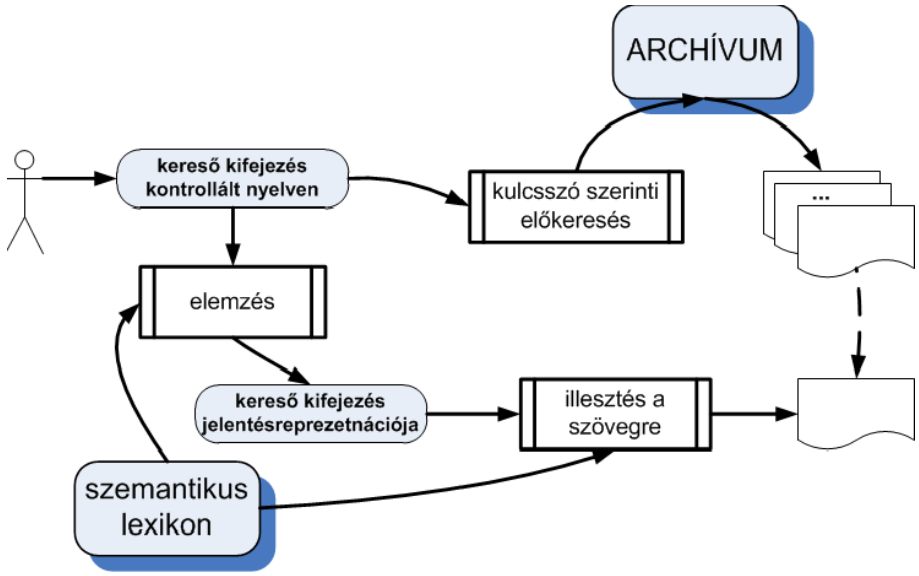
# MASZEKER: szemantikus keresőprogram

Hussami Péter[1]

[1]Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy J. u. 7
hussami@all.hu

A Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem Informatikai Tanszékcsoportja, valamint Könyvtár- és Humán Információtudományi Tanszéke közös projektet (TECH_08_A2/2-2008-0092) indított az Nemzeti Fejlesztési Ügynökség támogatásával. A projekt célja egy olyan, új elveken alapuló integrált keresőrendszer kifejlesztése, amely adaptált (statisztikai és szimbolikus alapú) technológiák és újszerű megoldások kombinálásán keresztül a keresést végző felhasználó szemantikai kompetenciáját az eddigieknél nagyobb mértékben kiaknázva teszi lehetővé a természetes nyelvi dokumentumtárakban (szövegekben) történő valóban tartalmi keresést. Egyszerűen szólva: a felhasználó jól formált frázisokkal, mondatokkal specifikálhatja, milyen tartalmú dokumentumokat keres.

A rendszer áttekintő architektúrája az 1. ábrán látható.




**1. ábra** A MASZEKER rendszer áttekintő architektúrája

Az ábrának megfelelően a releváns dokumentumok keresése a következő lépésekből áll:

1. a felhasználó egy kontrollált nyelven adja meg a keresőkifejezést,
2. szintaktikus és szemantikus elemzés előállítja keresőkifejezés jelentésreprezentációját,
3. szavak szerinti keresés előszűri az archívumot,
4. azokra a szövegszegmensekre, amelyekben a szavak szerinti keresés találatai vannak, illeszti a keresőkifejezés jelentésreprezentációját.

Az MSzNy VII konferencián tartott előadáson [1] ismertetésre kerültek a fenti elemek megvalósítására vonatkozó elméleti alapelvek, elsősorban a szemantikus reprezentáció felépítése mint sarokkő köré szervezve. Idén be kívánjuk mutatni a megvalósulás jelenlegi állapotát egy demó prezentálásával.

A demóban az archívumot szabadalmi leírások főigénypontjaiból összeállított dokumentumgyűjtemény alkotja[1]. A felhasználó a kontrollált nyelven megadhat keresőkifejezést. A keresőkifejezés több mondatból, ill. főnévi kifejezésből állhat, a megszorítások az egyértelműséget biztosítják – például korlátozzák az igeneves szerkezeteket. A felsorolásokat a felhasználónak jelölnie kell. A felhasználói interfész segíti a kontrollált nyelv szabályainak betartását, és a morfoszintaktikai elemzés eredménye alapján a rendszer ellenőrzi a szabályok betartását. A rendszer a keresőkifejezéshez illő frázisokat keres az igénypontok szövegében, és az eredményt a grafikus interfészen megmutatja, kiemelve azokat a szavakat, amelyekből álló frázist a keresőkifejezés egy szegmenséhez hasonlónak talált.

A végleges kiépítéshez képest a demó a következő egyszerűsítéseket alkalmazza:

- a kisméretű „archívum” miatt a kulcsszó szerinti előkeresés felesleges,
- a szemantikus lexikon kiépítettsége még messze van a kívánatostól, ezért a jelentésreprezentációk hiányosak lehetnek,
- a szintaktikus elemzés szemantikus kontrollja még nem teljes,
- a hasonlóság felismerésénél vannak figyelembe nem vett tényezők,
- a szabadalmi igénypontok szerkezetéből és a témakörből adódó heurisztikus megoldásokat kielégítően még nem alkalmaztuk[2],
- a relevancia meghatározása még nem eléggé kifinomult.

Mind a felismerés pontosságát, mind a performanciát a további kísérletek alapján javítani kívánjuk.

## Bibliográfia

1. Szőts M., Csirik J., Gergely T., Karvalics L:: MASZEKER: projekt szemantikus kereső technológia kidolgozására. In: Tanács A., Vincze V. (szerk.): MSzNy 2010 – VII Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 159–167

[1] A projekt egyik kiemelt felhasználási területe a szabadalmi keresés, s a demóban „gyógyhatású készítmények és kozmetikai szerek” témaköréből származó szabadalmakat használunk.

[2] Mind a szintaktikus, mind a szemantikus elemzést, mind a hasonlóság megállapítását nagyban befolyásolja, hogy milyen témakörben, milyen típusú dokumentumok közt keresünk.